Article

Journal of Educational and Behavioral Statistics
2023, Vol. 48, No. 4, pp. 420–453
DOI: 10.3102/10769986231155427

Is It Who You Are or Where You Are? Accounting for Compositional Differences in Cross-Site Treatment Effect Variation

Benjamin Lu
University of California, Berkeley

Eli Ben-Michael
Carnegie Mellon University

Avi Feller
University of California, Berkeley

Luke Miratrix
Harvard University

In multisite trials, learning about treatment effect variation across sites is critical for understanding where and for whom a program works. Unadjusted comparisons, however, capture "compositional" differences in the distributions of unit-level features as well as "contextual" differences in site-level features, including possible differences in program implementation. Our goal in this article is to adjust site-level estimates for differences in the distribution of observed unit-level features: If we can reweight (or "transport") each site to have a common distribution of observed unit-level covariates, the remaining treatment effect variation captures contextual and unobserved compositional differences across sites. This allows us to make apples-to-apples comparisons across sites, parceling out the amount of cross-site effect variation explained by systematic differences in populations served. In this article, we develop a framework for transporting effects using approximate balancing weights, where the weights are chosen to directly optimize unit-level covariate balance between each site and the common target distribution. We first develop our approach for the general setting of transporting the effect of a single-site trial. We then extend our method to multisite trials, assess its performance via simulation, and use it to analyze a series of multisite trials of adult education and vocational training programs. In our application, we find that distributional differences are potentially masking cross-site variation. Our method is available in the balancer R package.

Keywords: *multisite trials; generalizability; transportability; balancing weights; treatment effect variation*

1. Introduction

A central challenge for many questions in policy and the social sciences is to generalize (or "transport") results from a randomized control trial (RCT) to a target population (Egami & Hartman, 2021; Tipton, 2014). For instance, given an experimental evaluation of a job training program in one location, can we predict the program's effects in another location with different macroeconomic conditions and demographic composition (Hotz et al., 2005)?

In this article, we focus on generalizing or transporting effects in the context of multisite RCTs, where treatment assignment is randomized separately within each of several sites. Multisite RCTs have been used to study, for example, the effects of the Head Start program on childhood educational outcomes (Puma et al., 2010), of welfare-to-work programs on participant earnings (Kemple & Haimson, 1994; Riccio & Friedlander, 1992), of police body-worn camera usage on citizen complaints (Ariel et al., 2017), and of psychoeducational interventions on cancer patients' emotional health (Stanton et al., 2005).

Multisite RCTs show promise in part because they can reveal how treatment effects vary across different settings. Specifically, they can help disentangle treatment effect variation due to "compositional" differences in sites' distributions of observed unit-level features from variation due to other differences, such as "contextual" differences in site-level features or unobserved compositional differences (Rudolph et al., 2018). For instance, researchers might seek to understand the relationship between site-level impacts and site-level features, like the way the program was implemented in the site—after accounting for differences in observed unit-level features, like baseline income and education levels (see, e.g., Bloom et al., 2003; Bloom et al., 2020). Alternatively, researchers might focus on the extent to which observed unit-level features explain variation in site-level impacts (Weiss et al., 2014). There are a range of related quantities of interest in the literature (e.g., Bloom and Weiland, 2015; Bryk and Raudenbush, 1988; Djebbari and Smith, 2008; May et al., 2014; Raudenbush et al., 2012; Walters, 2015).

These inquiries can be interpreted statistically as special cases of transportability: If we transport the treatment effect from each site to the same target distribution of unit-level covariates, then we can attribute the change in cross-site treatment effect variation to those covariates. And we can attribute any remaining variation in the transported treatment effects to differences in site-level features and unobserved unit-level features.

Many transportation and generalization methods rely on weighting estimators, including doubly robust estimators that combine weighting and outcome modeling (see Egami & Hartman, 2020). Traditional inverse propensity score weighting (IPW) is the workhorse method (e.g., Rudolph et al., 2018). But it can perform poorly with many covariates or with extreme estimated propensity scores, and it requires unit-level data in the target distribution. More recently,

Josey, Yang, et al. (2020) instead propose using entropy balancing (Hainmueller, 2012), which finds weights that exactly balance a few covariates. But, while promising, entropy balancing is often infeasible with even a moderate number of covariates.

In this article, we develop a framework for transporting treatment effects using approximate balancing weights, a method recently developed in the observational causal inference literature (Ben-Michael, Feller, et al., 2021; Hirshberg and Wager, 2021; Zubizarreta, 2015). Our approach chooses weights to directly optimize unit-level covariate balance between each site and the target distribution. Unlike some other estimators, this approach can accommodate high-dimensional covariates, including higher-order interactions and kernels. And, in some cases, it can target an arbitrary covariate distribution without requiring unit-level data.

We first develop our approach for the general task of transporting the treatment effect from a single site; we view this as an important contribution in and of itself. We then adapt it to our motivating special case of decomposing treatment effect variation in multisite RCTs.[1] To decompose treatment effect variation, we first transport every site's treatment effect to a common target distribution of unit-level covariates. We can then descriptively analyze the cross-site variation in these transported treatment effects, net of differences in unit-level covariates. This problem formulation is quite flexible and does not require the strong multilevel linear modeling assumptions commonly used in these kinds of analyses (e.g., Bloom et al., 2003). At the same time, this more general formulation highlights that the underlying substantive questions are very difficult to address empirically: As we show in simulations, in all but the largest multisite trials, the sample size and number of sites are insufficient to draw meaningful conclusions (see also Weiss et al., 2017).

We apply our approach to an unusually large collection of seminal studies on welfare-to-work policies that offered job training and adult education to eligible people between 1988 and 1994. Specifically, we re-examine three separate multisite experiments: Project GAIN, Project Independence (PI), and the National Evaluation of Welfare-to-Work Strategies (NEWWS). Project GAIN provided basic education to those who needed remediation in math or language skills, as well as job-search assistance, unpaid work experience, and referrals for postsecondary education and vocational training. PI focused on low-cost job-search strategies and limited access to basic education. NEWWS consisted of six different programs focused on training and education. Together, these studies constitute a large multisite experiment, with 59 sites across seven states, totaling 69,399 participants. Within each site, participants were randomly assigned either to receive or to be barred from receiving the job training and adult education offered by the welfare-to-work program. The primary outcomes were employment status and earnings two years after random assignment. For each participant, we also observe 23 pretreatment covariates, including earnings prior to

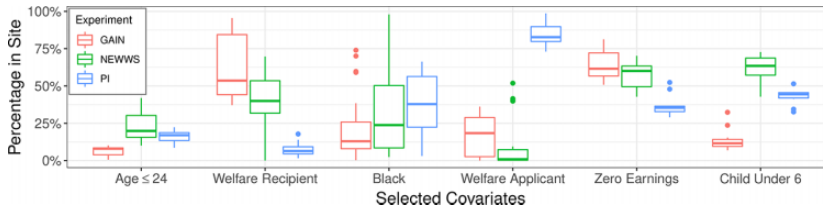

FIGURE 1. *Distributions of the marginal prevalence of selected binary covariates across the 59 sites of the multisite welfare-to-work experiments analyzed in this article. The covariates are distributed differently across sites. For example, each site in Project Independence had a greater proportion of welfare applicants than any site in the other two multisite experiments.*

randomization, number of dependent children, and high school completion. As Figure 1 shows, these covariates are distributed differently in each site and across the three experiments as a whole. As a result, direct comparisons of treatment effects across individual sites and the overall experiments might be difficult to interpret.

We therefore use our method to transport the treatment effects from the 59 sites to the same target covariate distribution. After reweighting, the estimated cross-site variation in average treatment effects (ATEs) is in fact larger for the transported estimates than for the unadjusted estimates. This suggests that differences in the sites' observed distributional makeup do not drive, and in fact could mask, differences in the overall efficacy of treatment at these sites. Moreover, while the unadjusted estimates suggest much larger impacts in GAIN sites than in PI sites, the adjusted estimates largely reverse this trend.

The remainder of this article is organized as follows. Section 2 establishes the basic setting for our problem, identifies our estimand, and distinguishes our work from recent literature. Section 3 introduces our proposed estimator based on approximate balancing weights. Section 4 shows how our setup and proposed estimators naturally extend to multisite RCTs and sketches our framework for decomposing treatment effect variation in this context. Section 5 compares our proposed estimator to other standard estimators via simulation. Section 6 applies our method to investigate treatment effect variation in the welfare-to-work experiments. Section 7 concludes. We provide an `R` package, `balancer`, to make our methods readily available to interested practitioners.

2. Setup

2.1. Estimand and Assumptions

We first consider a single-site RCT with a binary treatment, turning to multisite RCTs in Section 4. Of the n units in the experiment, n_1 are assigned to

treatment and n_0 are assigned to control. Let $Z_i \in \{0, 1\}$ be a binary treatment indicator for the ith unit. In addition to treatment assignment, we observe for each unit a vector of d pretreatment unit-level covariates X_i and the unit's posttreatment outcome Y_i. We adopt the potential outcomes framework and the stable unit treatment value assumption (i.e., there is no interference between units and only one form of treatment); see Neyman (1923) for their introduction and Rosenbaum (2009) for an overview. Under this framework, each unit has a potential outcome under treatment $Y_i(1)$ and a potential outcome under control $Y_i(0)$, with observed outcome $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$. Additionally, we assume that the units in the experiment are independently and identically distributed (i.i.d.) according to some distribution P: $\{X_i, Z_i, Y_i(0), Y_i(1)\} \overset{i.i.d.}{\sim} P$ for $i = 1, \ldots, n$.

Throughout this article, we use several conditional expectations defined over the experimental population's distribution P. First, we denote the propensity score for the experimental population as $e(x) := \mathbb{E}_P(Z|X = x)$. To simplify notation, we assume that this is some constant π known by design, but our results extend readily to nonconstant propensity scores. Second, we denote the prognostic score, the expected control potential outcome conditioned on the observed covariates, as $m_0(x) := \mathbb{E}_P\{Y(0)|X = x\}$. Analogously, we denote the expected potential outcome under treatment conditioned on the observed covariates as $m_1(x) := \mathbb{E}_P\{Y(1)|X = x\}$. Finally, we denote the conditional ATE (CATE) function as

$$\tau(\cdot) := \mathbb{E}_P\{Y(1) - Y(0)|X = x\}.$$

In terms of $\tau(\cdot)$, the ATE in our experimental population is

$$\tau := \mathbb{E}_P\{Y(1) - Y(0)\} = \mathbb{E}_P[E_P\{Y(1) - Y(0)|X\}] = \mathbb{E}_P\{\tau(X)\} = \int \tau(\cdot) dP(x). \quad (1)$$

We seek to characterize what the ATE in our site would be if it had some other distribution of subjects. Let P^* denote the distribution of some target population; the experimental population can but need not be a subset of this target population. Our primary estimand is then

$$\tau^* := \mathbb{E}_{P^*}[\mathbb{E}_P\{Y(1) - Y(0)|X\}] = \mathbb{E}_{P^*}\{\tau(X)\} = \int \tau(\cdot) dP^*(x). \quad (2)$$

The inner expectation in (2) is taken over the *experimental* population's distribution of potential outcomes conditioned on $X = x$—that is, it is simply our CATE function $\tau(\cdot)$. The outer expectation in Equation 2 is taken over the *target* population's distribution of observed covariates, P^*.

We make τ^* our estimand because it can help us understand treatment effect variation across different populations. Because the CATE function $\tau(x)$ in both τ and τ^* is the same, we can attribute any difference between τ and τ^* to differences in the observed covariate distributions, P and P^*. And, as we show when

we extend our setup to multisite RCTs in Section 4, analogous reasoning in the opposite direction—comparing ATEs defined by integrating different CATE functions over the same distribution of observed unit-level covariates—can help isolate differences in treatment effects due to site-level or unobserved unit-level covariate differences between sites.

In addition to the foundational assumptions stated at the start of this section, we make three assumptions that together identify τ^*. We assume that treatment is randomized, that the probability of treatment is strictly between 0 and 1, and that the target distribution of observed covariates overlaps with the experimental distribution of observed covariates.

Assumption 1 (Randomization): Each unit's treatment assignment is independent of its potential outcomes given observed unit-level covariates: $Y(0), Y(1) \perp\!\!\!\perp Z|X$ for $\{X, Z, Y(0), Y(1)\} \sim P$.

Assumption 2 (Positivity of Treatment Assignment): If x satisfies $p(x) > 0$, then $0 < e(x) < 1$.

Assumption 3 (Overlap): The target distribution of observed covariates is absolutely continuous with respect to the experimental distribution of observed covariates: $p^*(x) \ll p(x)$.

Assumption 3 requires that the target distribution of observed covariates have zero density (probability) everywhere that the experimental distribution of observed covariates has zero density (probability). This is analogous to the overlap condition commonly invoked in observational causal inference. Without it or some other modeling assumption, the CATE at covariate values observable in the target population but not in the experimental population cannot be identified.

One advantage of the framework we have established thus far is that the assumptions involved are fairly light. Randomization and positivity of treatment assignment are typically satisfied by design in RCTs. And the target population can be chosen to satisfy absolute continuity, although this can be more difficult in high dimensions (see, e.g., D'Amour et al., 2021, for discussion of issues that can arise). Under Assumptions 1–3, τ^* can be identified by

$$\begin{aligned} \tau^* &= \mathbb{E}_{P^*}\{\mathbb{E}_P(Y|Z=1,X) - \mathbb{E}_P(Y|Z=0,X)\} \\ &= \mathbb{E}_P\left[\left\{\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right\}\frac{dP^*}{dP}(X)\right], \end{aligned} \tag{3}$$

where $\frac{dP^*}{dP}(x)$ denotes the Radon–Nikodym derivative of the target distribution relative to the sample distribution, which generalizes the likelihood ratio of observing covariates x under the target and experimental distributions. Special cases include when all covariates are discrete, in which case $\frac{dP^*}{dP}(x)$ is the ratio of the probabilities of observing covariates $X = x$ in both distributions, and

when all covariates are continuous, in which case it is the ratio of density functions.

Throughout this article, we rely only on Assumptions 1 to 3 or, in Section 4, their extensions to multisite RCTs. However, we pause here to note that other work often invokes some form of the following assumption as well (Dahabreh et al., 2019).

Assumption 4 (Mean Generalizability of Treatment Effects): For all x satisfying $p^*(x) > 0$,

$$\mathbb{E}_{P^*}\{Y(1) - Y(0) | X = x\} = \mathbb{E}_P\{Y(1) - Y(0) | X = x\}.$$

Assumption 4 implies that the CATE function is the same in both the experimental population and the target population wherever the target covariate distribution has positive density. A common variation of this assumption is that units' potential outcomes are independent of the population to which they belong, conditional on covariates (Allcott, 2015; Flores & Mitnik, 2013; Hotz et al., 2005). Under Assumption 4, our estimand τ^* is equivalent to the ATE in the target population $\tilde{\tau}^* := \mathbb{E}_{P^*}\{Y(1) - Y(0)\} = \mathbb{E}_{P^*}[\mathbb{E}_{P^*}\{Y(1) - Y(0)|X\}]$. This is a standard estimand in the transportability and generalizability literature.

However, we do not use Assumption 4, nor do we make $\tilde{\tau}^*$ our estimand, because doing so would essentially assume away the phenomenon we seek to isolate. We seek to develop a realistic account of the extent to which observed unit-level covariate heterogeneity explains treatment effect variation across populations relative to other possible factors. But Assumption 4 assumes already that observed unit-level covariate heterogeneity explains all of it. Put differently, Assumption 4 enables externally valid generalization of the within-experiment ATE to the target population. Nie et al. (2021) assess sensitivity to this assumption for an estimation approach similar to ours.

2.2. Related Work

This article builds on a growing literature on the transportability and generalizability of treatment effects. For the most part, the setup and assumptions established in Section 2.1 are common in this literature (see, e.g., Allcott, 2015; Crepon et al., 2018; Dahabreh et al., 2019; Egami & Hartman, 2020; Hotz et al., 2005; Tipton, 2014).

However, two features of our work are worth emphasizing. First, we consider identification and estimation of site ATEs transported to an arbitrary target population. This target population can be finite or infinite. And it can but need not contain the experimental population. By contrast, previous work mostly focuses on transportation to an observed, finite target population of units and specifies whether the target population contains the experimental population or not (Dahabreh et al., 2020; Dahabreh et al., 2019; Hotz et al., 2005). Second, as

we note above, we seek to identify and estimate τ^* rather than $\tilde{\tau}^*$. So we do not require Assumption 4, which is often quite strong and difficult to verify in practice.

Despite these two differences, we can naturally adapt standard estimators for $\tilde{\tau}^*$ to estimate τ^* in our present setup. These estimators plug in estimates of the terms that identify τ^* in Equation 3. They generally mirror standard outcome-modeling and IPW estimators from the observational causal inference literature (see Ackerman et al., 2019; Egami & Hartman, 2020, for recent discussion). For example, if the conditional mean outcome model $m_z(x) := \mathbb{E}_P\{Y(z)|X = x\}$ is correctly specified for $z \in \{0, 1\}$, then the outcome-modeling estimator

$$\hat{\tau}^*_{\text{om}} := \mathbb{E}_{P^*}\{\hat{m}_1(X) - \hat{m}_0(X)\}, \tag{4}$$

is consistent for τ^* (see, e.g., Kern et al., 2016). Additionally, the IPW estimator

$$\hat{\tau}^*_{\text{ipw}} := \frac{1}{n}\sum_{i=1}^{n}\widehat{\frac{dP^*}{dP}}(X_i)\left\{\frac{Z_i}{\pi}Y_i - \frac{1 - Z_i}{1 - \pi}Y_i\right\}, \tag{5}$$

is consistent for τ^* if $\widehat{\frac{dP^*}{dP}}(x)$, the estimated ratio of the density of P^* to the density of P, is correctly specified. When both the experimental and target populations are finite, we can estimate $\widehat{\frac{dP^*}{dP}}(x)$ as follows. Let $E \in \{0, 1\}$ indicate inclusion in the experimental population and $T \in \{0, 1\}$ indicate inclusion in the target population. By Bayes' rule, the change in measure is

$$\frac{dP^*}{dP}(x) = \frac{p^*(x)}{p(x)} = \frac{\Pr(T = 1|X = x)}{\Pr(E = 1|X = x)}\frac{\Pr(E = 1)}{\Pr(T = 1)}. \tag{6}$$

We can then separately estimate the conditional probabilities in Equation 6. See Dahabreh et al. (2020) and Westreich et al. (2017) for analogous discussion in a similar setting. Finally, the doubly robust estimator

$$\begin{aligned}\hat{\tau}^*_{\text{dr}} := &\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{\frac{dP^*}{dP}}(X_i)\frac{Z_i}{\pi}\{Y_i - \hat{m}_1(X_i)\} + \mathbb{E}_{P^*}\{\hat{m}_1(X)\},\right] \\ &- \left[\frac{1}{n}\sum_{i=1}^{n}\widehat{\frac{dP^*}{dP}}(X_i)\frac{1 - Z_i}{1 - \pi}\{Y_i - \hat{m}_0(X_i)\} + \mathbb{E}_{P^*}\{\hat{m}_0(X)\}\right]\end{aligned} \tag{7}$$

is consistent for τ^* if either the outcome model or the change-of-measure model is correctly specified. Targeted maximum likelihood estimators for transported site effects (Rudolph & van der Laan, 2017) can similarly be adapted to our setup, but we do not discuss them at length.

IPW and outcome-modeling estimators, however, face well-known limitations. IPW might not achieve adequate finite-sample balance between the experimental and target populations (Ben-Michael, Feller, et al., 2021). This can occur when, for example, the conditional probability models are misspecified or when there is poor overlap. This problem can become particularly acute because the conditional probability estimates are inverted, as in Equation 6. Small errors in

estimating the conditional probabilities can produce large errors in the resulting effect estimate when inverted (Kang & Schafer, 2007). For this same reason, IPW can be unstable, with a few units given extremely large weights and thereby dominating the analysis (Robins et al., 2007). To be sure, these problems can arise in other weighting methods, including the one we propose using. But IPW offers only indirect, usually ad hoc remedies. Typically, an analyst who uses IPW iteratively fits different conditional probability models until one achieves adequate balance (Ben-Michael, Feller, et al., 2021; Harvey et al., 2017), with no clear guidance on what changes are best (Imbens & Wooldridge, 2009). The analyst may also trim or threshold the estimated probabilities afterward. See Crump et al., 2009; Ma & Wang, 2020; Yang & Ding, 2018, for recent advances in trimming IPW probability estimates.)

Like IPW, outcome modeling can depend heavily on model specification and the outsize influence of individual units (but see Kern et al., 2016, for evidence in favor of outcome modeling). While these vulnerabilities can be checked, their remedies are again usually indirect and ad hoc. Unlike IPW, outcome modeling is also prone to extrapolation: Estimates produced by outcome modeling are not necessarily sample-bounded in the sense of Robins et al. (2007) but instead can lie outside the support of the data (Chattopadhyay et al., 2020). Outcome models that rely on extrapolation can be sensitive to small changes in specification as a result, and this sensitivity is not typically reflected in the accompanying uncertainty quantification (Imbens, 2015; King & Zeng, 2006).

Our proposed weighting approach draws on the recent literature on approximate balancing weights in observational causal inference (see, e.g., Athey et al., 2018; Hirshberg & Wager, 2021; Zubizarreta, 2015). Ben-Michael, Feller, et al. (2021) give a recent review. These methods find weights that minimize a measure of covariate imbalance between two groups of units, typically between treated and control units in observational studies. A special case of this approach called entropy balancing (Hainmueller, 2012) arises when the weights can achieve *exact* balance in the covariates. In a set of papers closely aligned with ours, Josey, Berkowitz, et al. (2021) and Josey, Yang, et al. (2022) propose adapting entropy balancing to transport treatment effects. These papers offer an important conceptual advance relative to transportation methods that rely on traditional IPW, often producing better finite-sample performance and a smoother workflow. (See also Nie et al., 2021, who incorporate sensitivity analysis into a related framework.) However, weights that achieve exact balance are often infeasible in practice, even with a moderate number of covariates. Thus, our proposed approach extends their framework to allow for *approximate* balance, following a suggestion in the conclusion of Josey, Berkowitz, et al. (2021). Finally, for a related setting, Crepon et al. (2018) propose using outcome modeling for dimension reduction and then entropy balancing for transportation. It is

straightforward to adapt their approach to use approximate balancing weights, as proposed here, rather than entropy balancing.

3. Weighting Estimators for Transported Site ATEs

We propose estimating τ^* via a linear weighting estimator $\hat{\tau}^*$ with weights $\hat{\gamma} \in \mathbb{R}^n$:

$$\hat{\tau}^* := \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i \frac{Z_i}{\pi} Y_i - \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i \frac{1 - Z_i}{1 - \pi} Y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i \frac{Z_i - \pi}{\pi(1 - \pi)} Y_i. \tag{8}$$

Recall that we assume for clarity of exposition that the propensity score is constant: $e(X_i) = \pi$. This is satisfied by design in many simple experimental setups. The methods we propose in this section can be extended readily to non-constant propensity scores by substituting $e(X_i)$ for π.

We want to choose weights that optimize the performance of $\hat{\tau}^*$ as an estimator for τ^* by some metric. In this article, we specifically consider choosing weights to minimize the estimation error given by $\hat{\tau}^* - \tau^*$, adapting arguments and estimators from Athey et al. (2018) and Hirshberg and Wager (2021) designed for observational studies to estimate transported effects. We discuss in Section 3.1 one way of decomposing this estimation error. Then, we show in Section 3.2 how we can use convex optimization to choose weights that minimize in a controlled way the constituent parts of the decomposed estimation error.

3.1. Estimation Error Decomposition

First, we decompose the estimation error $\hat{\tau}^* - \tau^*$ into three terms: (1) covariate imbalance between treated and control units in the experimental population, (2) covariate imbalance between treated units in the experimental population and the target population, and (3) error due to noise. Doing so yields

$$\hat{\tau}^* - \tau^* = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i \frac{Z_i - \pi}{\pi(1 - \pi)} m_0(X_i)}_{\text{imbalance in } m_0(\cdot)} + \underbrace{\frac{1}{n\pi}\sum_{i=1}^{n} \hat{\gamma}_i Z_i \tau(X_i) - \mathbb{E}_{P^*}\{\tau(X)\}}_{\text{imbalance in } \tau(\cdot)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i \frac{Z_i - \pi}{\pi(1 - \pi)} \epsilon_i}_{\text{noise}}, \tag{9}$$

where $\epsilon_i := Y_i - m_0(X_i) - Z_i \tau(X_i)$. The first term in Equation 9 is the imbalance in the expected control potential outcome $m_0(\cdot)$ *within the study*. It is largely controlled by design via treatment randomization; if the weights $\hat{\gamma}$ are uniform, then this term will be zero in expectation by Assumption 1. However, any particular experiment will likely have some chance imbalance. In Section 3.2, we propose choosing weights $\hat{\gamma}$ that adjust for this in a manner similar to regression adjustment or poststratification (Lin, 2013; Miratrix et al., 2011). The second term in Equation 9 measures the discrepancy between the CATE averaged over our reweighted sample and our estimand. We are primarily interested in

controlling this term. The final term in Equation 9 arises from noise in the outcomes and is related to the variance of the estimator.

If we knew the true prognostic score $m_0(\cdot)$ and CATE $\tau(\cdot)$, then we could use Equation 9 as a guide to choose weights $\hat{\gamma}$: We would choose $\hat{\gamma}$ so that the imbalances in $m_0(\cdot)$ and $\tau(\cdot)$ are both zero or close to it. But we do not know the true prognostic score and CATE. So we posit a model class $\mathcal{M}$ for the prognostic score $m_0(\cdot)$ and a model class $\mathcal{T}$ for the CATE $\tau(\cdot)$. An example of a model class is the set of linear models $\{\beta_0 \cdot x | \beta_0 \in \mathbb{R}^d\}$, which corresponds to assuming that the prognostic score (or CATE) is linear in the covariates. We consider two main sets of model classes in Section 3.3. With these model classes, we can bound the estimation error by

$$|\hat{\tau}^* - \tau^*| \leq \sup_{m \in \mathcal{M}} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}_i \frac{Z_i - \pi}{\pi(1-\pi)} m(X_i) \right| + \sup_{\tau \in \mathcal{T}} \left| \frac{1}{n\pi} \sum_{i=1}^{n} \hat{\gamma}_i Z_i \tau(X_i) - \mathbb{E}_{P^*}\{\tau(X)\} \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}_i \frac{Z_i - \pi}{\pi(1-\pi)} \epsilon_i \right|. \tag{10}$$

This replaces the imbalance in the true prognostic score and CATE with the *worst-case* imbalance across all potential prognostic scores $m \in \mathcal{M}$ and CATEs $\tau \in \mathcal{T}$. Doing so gives us a feasible guide for choosing the weights: If we can place modeling restrictions on the prognostic score and the CATE, then we can try to control the components of Equation 10. We turn to this next.

3.2. Minimizing the Worst-Case Estimation Error

To control the estimation error (Equation 9) given model classes for the prognostic score $\mathcal{M}$ and the CATE $\mathcal{T}$, we use convex optimization to find weights that solve

$$\begin{aligned} \min_{\gamma} \quad & \sup_{m \in \mathcal{M}} \left| \frac{1}{n} \sum_{i=1}^{n} \gamma_i \frac{Z_i - \pi}{\pi(1-\pi)} m(X_i) \right|^2 \\ & + \sup_{\tau \in \mathcal{T}} \left| \frac{1}{n\pi} \sum_{i=1}^{n} \gamma_i Z_i \tau(X_i) - \mathbb{E}_{P^*}\{\tau(X)\} \right|^2 + \lambda \sum_{i=1}^{n} \gamma_i^2 \left(\frac{Z_i}{\pi} + \frac{1 - Z_i}{1-\pi} \right) \\ \text{subject to} \quad & \sum_{i=1}^{n} Z_i \gamma_i = n_1, \quad \sum_{i=1}^{n} (1 - Z_i) \gamma_i = n_0, \quad \gamma_i \geq 0. \end{aligned} \tag{11}$$

The objective in this optimization problem (Equation 11) directly targets the upper bound on the estimation error (Equation 10). First, it optimizes external validity by weighting treated units so that the weighted average of those units' CATEs resembles the target distribution's ATE τ^*. It does this for the worst-case CATE function. At the same time, it maintains internal validity by weighting the control units so that they are comparable to the weighted distribution of treated units

with respect to the prognostic score. The final term in the objective penalizes the weights for nonuniformity via an L^2 regularization term; this is a proxy for the variance due to noise in Equation 10. We include a regularization hyperparameter λ that controls this trade-off between better balance (lower bias) and more uniform weights (lower variance). We discuss the choice of λ in practice in Section 6.

The optimization problem (Equation 11) includes three constraints that stabilize the estimator. The first two constrain the weights on the treated units and the control units to sum to the number of treated units n_1 and the number of control units n_0, respectively. This ensures that the estimator is robust to constant shifts in the outcome (Ben-Michael, Feller, et al., 2021). The third constraint restricts the weights to be nonnegative. This prohibits extrapolation from the support of the experimental sample when estimating τ^* (see Ben-Michael, Feller, & Rothstein, 2021; King and Zeng, 2006; Zubizarreta, 2015). Taken together, these constraints also ensure that the estimator is sample-bounded in the sense that the resulting estimate is a convex combination of observed responses (see Robins et al., 2007).

To estimate the standard error of $\hat{\tau}$, we first estimate the conditional expected potential outcome functions $\hat{m}_1(\cdot)$ and $\hat{m}_0(\cdot)$ via regularized weighted least squares using the weights $\hat{\gamma}$. Then, we compute the squared standard error as

$$\hat{V} = \frac{1}{n_1^2}\sum_{i=1}^{n} Z_i \hat{\gamma}_i^2 (Y_i - \hat{m}_1(X_i))^2 + \frac{1}{n_0^2}\sum_{i=1}^{n} (1 - Z_i)\hat{\gamma}_i^2 (Y_i - \hat{m}_0(X_i))^2. \tag{12}$$

The above expression for $\hat{V}$ depends on the constraints on the sum of the weights (the first two constraints in Equation 11).

Hirshberg et al. (2019) and Ben-Michael, Feller, et al. (2021) give technical conditions for $\hat{\tau}^*$ to be asymptotically normal around the true value τ^* and for Equation 12 to give standard error estimates that are consistent. Following our discussion on implementation below, we estimate the two potential outcome models $\hat{m}_1(\cdot)$ and $\hat{m}_0(\cdot)$ via ridge regression.

3.3. Implementation

We implement the optimization problem in Equation 11 for two pairs of model classes when the target distribution P^* is an empirical distribution over m units with covariates $\{\widetilde{X}_1, \ldots, \widetilde{X}_m\}$, where the tildes emphasize that the target covariates are potentially different from the experimental sample. In the first, both $m_0(\cdot)$ and $\tau(\cdot)$ are linear in transformations of the observed covariates. That is, $\mathcal{M} = \{\beta_0 \cdot \phi_0(x) | \parallel \beta_0 \parallel \leq C_0\}$ and $\mathcal{T} = \{\beta_\tau \cdot \phi_\tau(x) | \parallel \beta_\tau \parallel \leq C_\tau\}$, where $\phi_0(\cdot)$ and $\phi_\tau(\cdot)$ are some specified transformations of the covariates and C_0 and C_τ are nonnegative constants. In this formulation, we allow the models to use different transformations of the covariates $\phi_0(\cdot)$ and $\phi_\tau(\cdot)$ so that, for example, $m_0(\cdot)$ can have more expressive transformations that create a more flexible basis expansion while $\tau(\cdot)$ can be restricted to a simpler CATE function (see Hahn et al., 2020; Künzel et al., 2019, for related discussion). By the Cauchy–Schwarz

inequality, the special case of the optimization problem in Equation 11 when we use the L^2 norm $\|\cdot\|_2$—that is, when we seek to minimize the worst-case sum of squared imbalances—is[2]

$$\min_{\gamma} \left\| \frac{1}{n}\sum_{i=1}^{n}\gamma_i \frac{Z_i-\pi}{\pi(1-\pi)}\phi_0(X_i) \right\|_2^2 + \left\| \frac{1}{n\pi}\sum_{i=1}^{n}\gamma_i Z_i \phi_\tau(X_i) - \frac{1}{m}\sum_{\ell=1}^{m}\phi_\tau(\widetilde{X}_\ell) \right\|_2^2 + \lambda\sum_{i=1}^{n}\gamma_i^2\left(\frac{Z_i}{\pi}+\frac{1-Z_i}{1-\pi}\right)$$
$$\text{subject to } \sum_{i=1}^{n} Z_i\gamma_i = n_1, \quad \sum_{i=1}^{n}(1-Z_i)\gamma_i = n_0, \quad \gamma_i \geq 0. \tag{13}$$

This optimization problem is a quadratic program (QP), so we can efficiently solve it even with many units and high-dimensional $\phi_0(\cdot)$ and $\phi_\tau(\cdot)$ by, for example, the alternating direction method of multipliers (Boyd et al., 2010). Reweighting to the target distribution requires only the *average* of $\phi_\tau(\widetilde{X}_\ell)$ over the target distribution. Thus, we can solve Equation 13 without knowing the full underlying distribution $\{\widetilde{X}_1, \ldots, \widetilde{X}_m\}$. This reduced data requirement can be useful in practice.

Second, we can generalize this setup using kernels; for a review, see Ben-Michael, Feller, et al. (2021). Formally, let $\mathcal{H}_{k_0}$ and $\mathcal{H}_{k_\tau}$ be reproducing kernel Hilbert spaces (RKHS) with associated kernels $k_0 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $k_\tau : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, where, recall, d is the dimension of X. Informally, these kernels k_0 and k_τ calculate "similarity scores" between points in the covariate space, with potentially different measures of similarity for the two kernels. Then, we can consider the model classes for the prognostic score and the CATE to be $\mathcal{M} = \{m_0(\cdot) \in \mathcal{H}_{k_0} | \; \| m_0 \|_{\mathcal{H}_{k_0}} \leq 1\}$ and $\mathcal{T} = \{\tau(\cdot) \in \mathcal{H}_{k_\tau} | \; \| \tau \|_{\mathcal{H}_{k_\tau}} \leq 1\}$, respectively, where $\|\cdot\|_{\mathcal{H}_k}$ is the RKHS norm induced by kernel k. By the reproducing property, we can write these model classes as linear in infinite-dimensional bases, $\phi_0(\cdot)$ and $\phi_\tau(\cdot)$, defined by the kernel functions k_0 and k_τ, respectively.

The resulting specialization of Equation 11 involves the imbalances in the *infinite*-dimensional transformations of the covariates, $\phi_0(X_i)$ and $\phi_\tau(X_i)$. However, we do not need to explicitly compute the imbalance in these infinite dimensions. Instead, we can use the "kernel trick"—$\langle \phi_0(x), \phi_0(y) \rangle = k_0(x,y)$ and $\langle \phi_\tau(x), \phi_\tau(y) \rangle = k_\tau(x,y)$—and write the balancing weights optimization problem as

$$\begin{aligned} \min_{\gamma} \quad & \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_i\gamma_j \frac{Z_i-\pi}{\pi(1-\pi)}\frac{Z_j-\pi}{\pi(1-\pi)}k_0(X_i,X_j) \\ & + \frac{1}{n^2\pi^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_i\gamma_j Z_i Z_j k_\tau(X_i,X_j) - 2\frac{1}{nm\pi}\sum_{i=1}^{n}\sum_{\ell=1}^{m}\gamma_i Z_i k_\tau(X_i,\widetilde{X}_\ell) \\ & + \lambda\sum_{i=1}^{n}\gamma_i^2\left(\frac{Z_i}{\pi}+\frac{1-Z_i}{1-\pi}\right) \\ \text{subject to} \quad & \sum_{i=1}^{n} Z_i\gamma_i = n_1, \quad \sum_{i=1}^{n}(1-Z_i)\gamma_i = n_0, \quad \gamma_i \geq 0. \end{aligned} \tag{14}$$

As above, this optimization problem is a QP.[3] To compute the imbalance terms, we only need to compute the kernel evaluations $k_0(X_i, X_j)$ and $k_\tau(X_i, X_j)$ for each pair of units in the experimental population, as well as the expected kernel evaluation in the target distribution $\mathbb{E}_{P^*}\{k_\tau(X_i, X)\}$ for each treated experimental unit X_i.

The kernel functions are the key determinants of this second approach. From one view, they define the model classes for $m_0(\cdot)$ and $\tau(\cdot)$. From another view, they define the transformations of the covariates we seek to balance. While this is a nonparametric approach, it is not a panacea and can still be subject to model misspecification. In our simulations in Section 5, we consider the commonly used radial basis function (RBF) kernel $k(x, x') = \exp(-\|x - x'\|^2)$ (see, e.g., Hastie et al., 2009, § 5.8 and § 12.3).[4] Although popular, this choice of kernel still makes fairly strong assumptions on the model class, namely that the models are infinitely differentiable and hence very smooth. Thus, this kernel can lead to model misspecification and get the error bound in Equation 10 wrong if, for example, the true model is not so smooth. In addition, the kernel-based approach has a higher data requirement than the finite-basis-expansion approach: Unlike in Equation 13, the full set of unit-level covariates in the target population $\{\widetilde{X}_1, \ldots, \widetilde{X}_m\}$ is needed in Equation 14 to compute the kernel evaluations between the treated units and the target population, $k_\tau(X_i, \widetilde{X}_\ell)$.

4. Extension to Multisite RCTs

In this section, we extend our proposal to multisite RCTs. We begin with the formal setup and then turn to substantive questions we can address using the adjusted estimates.

4.1. Setup

Consider a multisite RCT with n total units across J sites, in which the n_j units in site j are separately randomized with probability of treatment π_j. To extend our framework to this setting, we simply repeat the analysis in Sections 2 and 3 separately for each of the J sites. Our estimand for the jth site is the ATE in the site if its population had the same distribution of observed unit-level covariates as the target population:

$$\tau_j^* := \mathbb{E}_{P^*}[\mathbb{E}_P\{Y(1) - Y(0)|X, S = j\}] = \mathbb{E}_{P^*}\left\{m_{1j}(X) - m_{0j}(X)\right\} = \int \tau_j(x)\, dP^*(x), \tag{15}$$

where $S \in \{1, \ldots, J\}$ indicates the unit's site membership, $m_{zj}(x) := \mathbb{E}_P\{Y(z)|X = x, S = j\}$ is the conditional mean potential outcome function in the jth site, and $\tau_j(x) = m_{1j}(x) - m_{0j}(x)$ is the CATE in the jth site. If Assumptions 1–3 hold within each site, then τ_j^* is identifiable as in Equation 3 for each $j \in \{1, \ldots, J\}$.

Our weighting approach to estimating τ^* in the single-site setting can be directly applied to estimate τ_j^* for any $j \in \{1, \ldots, J\}$: We simply treat the units in the *j*th site as the experimental population and ignore units in other sites. We thus refer to our weighting estimator in this multisite context as $\hat{\tau}_j^*$. By sequentially solving Equation 11 for each $j \in \{1, \ldots, J\}$, we can obtain the weights $\hat{\gamma}$ that define our weighting estimators $\hat{\tau}_1^*, \ldots, \hat{\tau}_J^*$. The standard outcome-modeling, IPW, and doubly robust estimators from Section 2.2 can similarly be extended to this multisite setting by restricting attention only to the units from the site of interest *j*.

4.2. Using the Adjusted Estimates

We now turn to the substantive questions we can address given the unadjusted site-level ATEs $\tau_1, \ldots, \tau_J$ and the adjusted site-level ATEs $\tau_1^*, \ldots, \tau_J^*$. A large and diverse literature seeks to understand how compositional and contextual differences across sites contribute to treatment effect variation; examples include Dehejia (2003), Bloom et al. (2003), Bloom and Weiland (2015), Bryk and Raudenbush (1988), Djebbari and Smith (2008), May et al. (2014), Raudenbush et al. (2012), Weiss et al. (2014), Walters (2015), and Crepon et al. (2018). We focus on three main sets of inquiries: (1) comparing adjusted estimates across sites, (2) examining the distribution of adjusted estimates, and (3) comparing adjusted and unadjusted estimates.

4.2.1 Comparing adjusted estimates across sites. As we discuss above, the most immediate application of this framework is to compare adjusted estimates across sites, often with the goal of identifying better-performing programs; prominent examples in the context of job training and adult education include Bloom et al. (2003), Hotz et al. (2005), Dehejia (2003), and Crepon et al. (2018). Comparing unadjusted impacts between two sites captures a myriad of differences, including differences in the populations served. However, adjusting for observable unit-level differences allows for more direct apples-to-apples comparisons between sites, even if many other differences remain. In particular, using the adjusted site-level estimates to identify sites of interest avoids inadvertently identifying sites based on their idiosyncratic (observed) population characteristics. For example, in our results in Section 6, the unadjusted estimates suggest much larger impacts in GAIN sites than in PI sites, but the adjusted estimates largely reverse this trend.

Finally, while we do not explore this here, the adjusted site-level estimates can also be inputs into site-level regressions and other models that seek to explain differences across sites. Using the adjusted site-level estimates removes variation due to distributional differences in observed unit-level covariates and thus reduces the possibility of reporting spurious relationships between site-level covariates and the site ATEs.

4.2.2. Understanding the distribution of adjusted estimates. A more general goal is to understand the cross-site distribution in the adjusted estimates, which has important implications for policy and research (see Weiss et al., 2017, for a thorough discussion of cross-site variation).

One measure of the distribution is the variance of the ATEs across sites. Let θ be the variance of $\tau_1, \ldots, \tau_J$, and let θ^* be the variance of $\tau_1^*, \ldots, \tau_J^*$. We can then compare θ and θ^* to understand how individual characteristics relate to cross-site differences in aggregate. For example, even if a program has a positive estimated effect overall, substantial cross-site variation introduces policy risk for local decision-makers. This is especially true if such variation is difficult to predict. Understanding θ^*, the variation in *adjusted* site-level impacts, therefore gives policymakers a sense of this risk. θ^* is also important for research design and assessing generalizability: If there is little variation in the adjusted estimates, then researchers can be more confident in generalizing from the purposive sample to new sites (see Tipton, 2014). See Online Appendix A for how we estimate θ and θ^* using meta-analytic techniques.

4.2.3. Comparing adjusted and unadjusted estimates. Finally, we can compare the adjusted and unadjusted site-level estimates to quantify the level of cross-site variation explained by distributional differences in unit-level covariates. We can view this as the natural multisite extension to the literature on decomposing treatment effect variation in randomized trials (see, e.g., Ding et al., 2019; Djebbari & Smith, 2008; Schochet et al., 2014).

We can directly compare θ and θ^*, the cross-site variances of the unadjusted and adjusted estimates, respectively, with a pseudo-R^2 measure of

$$R^2 := 1 - \frac{\theta^*}{\theta}. \tag{16}$$

This R^2 measure approximates the amount of site-level variability attributable to observed unit-level covariates. This is only a pseudo-R^2 measure since variability might actually increase after transportation ($\theta^* > \theta$) if differences in observed unit-level covariate distributions mask variability due to differences in site-level features. An increase in variability could be a noteworthy finding. In fact, we find such an increase in our empirical application.

The R^2, θ, and θ^* values can help inform the design of future studies, especially when selecting representative sites for a generalizable experiment (Tipton, 2014). In particular, if unit-level covariates explain very little of the cross-site variation, then taking those covariates into account when selecting sites is less imperative (see Tipton et al., 2019).

We caution that comparisons of τ_j to τ_j^* reveal the influence of sites' *distributions* of observed unit-level covariates on the *sites' ATEs*. They do not necessarily reveal the influence of the covariates themselves on *individual treatment effects*, as a standard analysis of heterogeneous treatment effects in a single-site

study might. For example, τ_j and τ_j^* can be similar even when some unit-level covariates strongly moderate the treatment effect for individual units. And they can be different even though all the unit-level covariates only weakly influence the treatment effect for individual units. For intuition on this, consider a linear treatment effect model for site j where the CATE is linear in observed unit-level covariates with coefficient vector β_j. We can decompose the difference between τ_j and τ_j^* as

$$\tau_j - \tau_j^* = \beta_j \cdot \{\mathbb{E}(X|S = j) - \mathbb{E}_{P^*}(X)\}.$$

This expression shows that the magnitude of $\tau_j - \tau_j^*$ depends on several factors. Even if the magnitudes of β_j are large, the difference $\tau_j - \tau_j^*$ can be small if these coefficients effectively cancel each other out or if the differences in the sites' covariate averages are small. Conversely, if the magnitudes of β_j are small, the difference between estimates can still be large if there are substantial differences in the sites' covariate averages.

5. Simulation Study

In this section, we examine the behavior of our proposed estimators and compare them to outcome-modeling, IPW, and doubly robust estimators via simulation. We base our simulations on data from the welfare-to-work experiments analyzed by Bloom et al. (2003), introduced in Section 1. In these experiments, treated units were given adult education and job training through the welfare-to-work program, while untreated units were not. Our outcome of interest is log earnings in dollars; for units with zero earnings, we add US$100 before logging to avoid undefined outcomes.

We generate data for our simulations using this dataset as follows. First, we define the CATE $\tau_j(\cdot)$ in the jth site to be a linear function of a subset of our unit-level covariates X_i—prior log earnings; indicators for whether prior earnings were zero, between zero and US$2,500, between US$2,500 and US$7,500, or greater than US$7,500; indicators for being less than 25 years old, between 25 and 34 years old, between 35 and 44 years old, or greater than 44 years old; an indicator for having a child younger than 6 years old; an indicator for applying for welfare; and an indicator for receiving welfare continuously for the past year—with an intercept term specific to the multisite experiment to which the jth site belongs. We base the coefficients in the CATE function on the coefficients on the interactions between treatment and covariates in a linear regression of the outcome on treatment and covariates in the actual experimental data. See Online Appendix B for the precise coefficient values and additional details.

In each of the 100 simulation repetitions, we generate a bootstrap sample of each of the 59 sites, define the potential outcomes under control $Y_i(0)$ to be the observed outcomes of the bootstrapped units plus a noise term

$\epsilon_i \sim^{i.i.d.} \mathcal{N}(0, 0.5^2)$, and generate potential outcomes under treatment $Y_i(1)$ by adding $\tau_j(X_i)$ to $Y_i(0)$. We then randomize treatment assignment, keeping the proportion of treated units in each site the same as in the original dataset, and set the observed outcome to the corresponding potential outcome.

In each repetition, our estimands are the site ATEs transported to the overall bootstrapped population. In other words, we seek to estimate $\tau_1^*, \ldots, \tau_{59}^*$ defined in Equation 15, where P^* is the bootstrapped population of all units across all three experiments. Note that, for any two sites k and ℓ, τ_k^* and τ_ℓ^* differ only by their experiment-specific intercepts since the unit-level covariates define $\tau_j(\cdot)$ in the same way for all $j \in \{1, \ldots, 59\}$ by design.

We estimate $\tau_1^*, \ldots, \tau_{59}^*$ using two versions of our weighting estimator $\hat{\tau}_j^*$. The first, which we refer to as "linear," optimizes the weights over the class of linear functions of X_i that could define $m_{0j}(\cdot)$ and $\tau_j(\cdot)$, as in Equation 13. The second version, which we refer to as "RBF," optimizes the weights to solve Equation 14, where the kernel function for the control potential outcomes, $k_0(\cdot, \cdot)$, is the RBF kernel $k(x, x') = \exp(-\|x - x'\|^2)$, and the kernel for the CATE, $k_\tau(\cdot, \cdot)$, is the linear kernel $k(x, x') = x \cdot x'$; this latter kernel corresponds to a standard linear basis, such as that implied by ridge regression. Both the linear and RBF versions of $\hat{\tau}_j^*$ require only the covariate means, not the full covariate distribution, of the target population because they assume the CATE is linear in covariates. In this simulation, both versions correctly specify $\tau_j(\cdot)$, but the RBF version of $\hat{\tau}_j^*$ allows a more flexible model for $m_{0j}(\cdot)$ than the linear version does. Since only one covariate is continuous, however, it is not obvious that this greater flexibility will appreciably advantage the RBF version.

For comparison, we also estimate $\tau_1^*, \ldots, \tau_{59}^*$ using the outcome-modeling, IPW, and doubly robust estimators described in Section 2.2. For these models, we use logistic regression to estimate $\Pr(S = j|X = \mathrm{x})$ and linear regression to estimate $m_{zj}(\cdot)$. Finally, as a benchmark, we also compute the naive, unadjusted difference-in-sample-means estimate of each site ATE, $\bar{Y}_{j1} - \bar{Y}_{j0}$ for $j \in 1, \ldots, 59$, to estimate $\tau_1^*, \ldots, \tau_{59}^*$.

We ran the simulation as described above three times. Each time, we changed the magnitude of the coefficients on the unit-level covariates defining the CATE. Specifically, we ran the simulation with the coefficients multiplied by 1, 5, and 10. This range of magnitudes reflects the varying values the pseudo-R^2 of Equation 16 might take in a multisite experiment—that is, the varying degrees to which distributional differences in observed unit-level covariates could explain ATE heterogeneity across sites relative to other factors. For context, Table 1 provides summary statistics of the data-generating processes used in the three versions of the simulation.

We evaluate the estimators' performance by computing the root mean squared error (RMSE) and bias for each of the 59 sites over the simulation repetitions.

TABLE 1.
Summary Statistics of the Data-Generating Processes Used in the Simulations

CATE Multiplier	R^2	Range of $\tau_j(x)$	*SD* of *Y*	Range of τ_j^*	Avg $Y(0)$	Avg $Y(1)/Y(0)$ ($)
1	.01	$(-0.09, 2.01)$	2.29	$(0.34, 1.75)$	6,380	2.55
5	.18	$(-0.42, 4.42)$	2.45	$(1.70, 3.11)$	6,380	11.88
10	.46	$(-0.84, 7.44)$	2.96	$(3.41, 4.81)$	6,380	107.73

Note. We perform the same simulation three times, each with a different multiplier on the coefficients of the unit-level covariates that define the CATE. CATE = conditional average treatment effect.

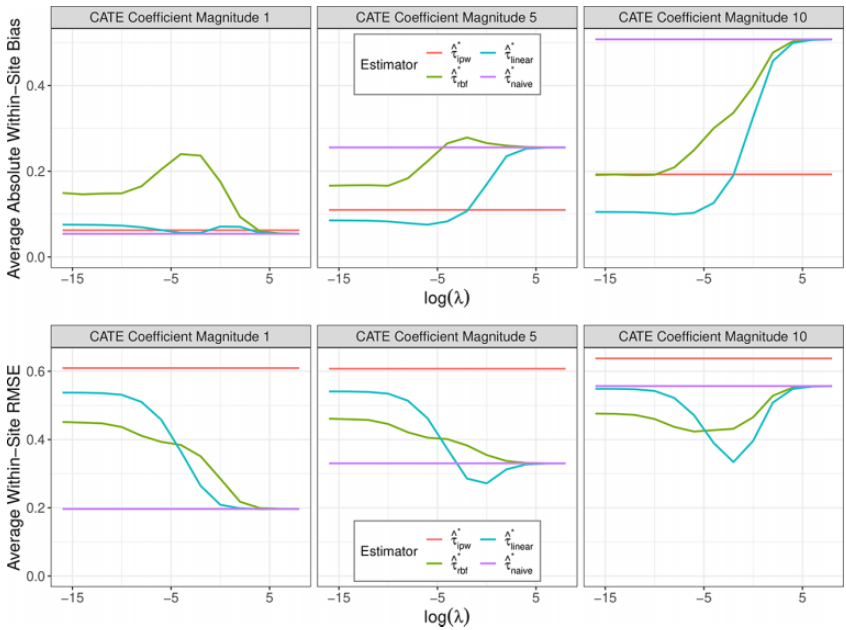

FIGURE 2. *Mean absolute bias (top) and root mean squared error (bottom) of each estimator over 100 repetitions of the welfare-to-work simulation. The conditional average treatment effect is defined as a sparse linear function of unit-level covariates plus an experiment-specific intercept term. Results for the outcome-modeling and doubly robust estimators are omitted because they are substantially larger; see Online Appendix C for those results. Table 1 offers summary statistics to contextualize these results.*

Figure 2 shows the RMSE and mean absolute bias of each estimator, averaged over the 59 sites. We show the RMSE and absolute bias of the outcome-modeling and doubly robust estimators separately in Figure C.1 of Online Appendix C because they are substantially larger than those of the estimators shown here.

Based on our examination of individual model fits, we determined that this was because one or both of the outcome models tends to have unstable coefficient estimates in a handful of sites, likely due to limited sampling and multicollinearity. In such cases, the outcome model struggles to stably extrapolate, as discussed in Section 2.2, and produces extreme predicted outcomes for some units. See Online Appendix C for details).

Although we logged the outcome, we can conceptualize these performance metrics in terms of dollars. For example, the exponentiated bias represents how disproportionate the estimated ratio between the treated and control potential outcomes (in dollars) is from the true ratio. Thus, the naive estimator's bias of about 0.25 in log earnings on average across sites in the middle panel of Figure 2 indicates that its estimate of the ratio between treated and control potential outcomes (in dollars) was about $e^{0.25} \approx 1.28$ times the true ratio in the site after adjustment. Table 1 lists the average control potential outcomes and the average ratio between the treated and control potential outcomes for each simulation setup. From those values, we see that the naive estimator's bias of about 0.25 in log earnings corresponds to a positive bias of about US$21,000 over the true difference of about US$69,000.

The right panels of Figure 2, where unit-level covariates explain a substantial amount of cross-site ATE heterogeneity, illustrate the bias-variance trade-off of regularizing our weighting estimators. Three trends in the behavior of our weighting estimators are apparent there. First, our weighting estimators are least biased when almost completely unregularized. Second, their RMSEs are lowest at some medium level of regularization. Third, they reduce to the naive estimator when heavily regularized.

The different simulation setups also show how different estimators compare. In the left panels, where unit-level covariates explain little heterogeneity, adjustment is unnecessary. So the naive estimator has lower bias and RMSE than the other estimators, which are essentially overfitting to noise. In the middle panels, where unit-level covariates explain more heterogeneity, IPW does as well as our weighting estimators at reducing bias without requiring a choice of regularization, although it comes at the cost of higher RMSE. In the right panels, where unit-level covariates explain the most heterogeneity, the linear weighting approach can achieve lower bias than IPW, with a much lower overall RMSE. These results indicate that the inverse propensity weights are relatively unstable compared to the balancing weights in these simulations due to overlap issues across sites, leading the IPW estimator to have higher variance.

In this simulation, we find that using the RBF kernel for the prognostic score has higher bias and RMSE than the linear approach that makes the more restrictive assumption that the prognostic score is linear, except in very low regularization settings where the RMSE is poor regardless. Practically, this indicates that controlling the imbalance in the covariate means between the treated and control groups is sufficient to control the bias. In attempting to

control a flexible, nonparametric measure of balance instead, the RBF approach is failing to control balance in the marginal covariates, leading to increased bias. By the same token, the RBF weights are made more extreme to control this nonparametric balance, leading to a higher-variance estimator without the benefits of reduced bias. Based on these results, we use the linear approach in our empirical analysis in Section 6.

To better understand how the estimators perform under a wider variety of settings, we rerun these simulations with model misspecification and with reduced resampling proportions. Details and results are in Online Appendix C. Overall, we find that the performance of the IPW and balancing-weights estimators predictably deteriorates but remains fairly robust. We also find that the same trends in relative performance discussed here largely hold.

5.1. Results With Smaller Site and Sample Sizes

Our main application and corresponding calibrated simulation study have 69,399 units across 59 sites. Few multisite experiments are this large. To understand how the estimators discussed in this article perform in smaller multisite experiments, we rerun the simulations above twice. In one, we resample only 34% of the sites (20 sites) and 20% of the units (13,880 units) in each repetition, leading to roughly 700 units per site on average. In the other, we resample only 25% of the sites (15 sites) and 5% of the units (3,470 units), leading to about 230 units per site on average. We keep the relative proportion of units in each site the same as in the full data set.

Appendix Figures C.3 and C.4 show the resulting mean absolute bias and RMSE of the estimators for these two simulation studies. Results for the outcome-modeling and doubly robust estimators are omitted because they are substantially larger. First, the reduced sample size worsens the performance of all estimators as expected. But as sample size goes down, the benefit of weighting relative to just using the naive estimator goes down substantially as well. With smaller sample sizes, the weights cannot achieve as much balance, so they cannot reduce bias as effectively. Thus, consistent with other research on multisite trials (Bloom & Spybrook, 2017; Hedges & Pigott, 2001; Raudenbush & Liu, 2000), we find that our ability to extract information on cross-site distributions decreases as the size and number of sites decrease. Unlike these existing results for untransported estimates, however, we must also consider the *effective* sample size after transportation. This is typically smaller than the nominal sample size, further reducing our power.

6. Empirical Application

We now apply our method to characterize treatment effect variation in the welfare-to-work experimental data. Recall that treated subjects were offered job training and adult education through the welfare-to-work program, while untreated subjects were not. Our outcome of interest in this analysis is a binary indicator of employment at

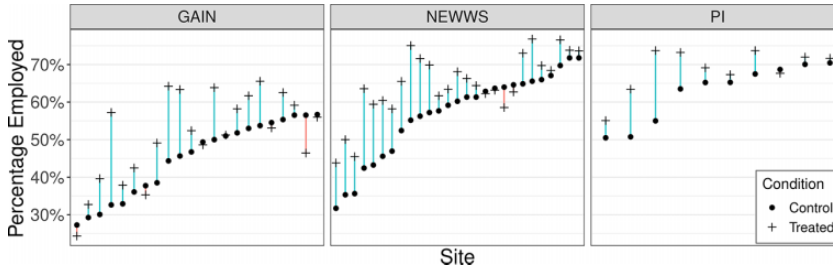

FIGURE 3. *Employment rate in the treated and control group of each site two years after randomization. Sites where the treated group has a higher employment rate than the control group are shown in blue.*

any point over the two years after treatment assignment, measured by nonzero total earnings over those two years. Figure 3 plots the average outcome in each treatment condition in each site. In most sites, the employment rate among treated units is higher than among control units; estimated impacts are the differences between the "+"s and the dots. We also see substantial heterogeneity in terms of baseline employment rates, and that the sites in the PI study generally have higher rates of employment. Pooling all sites, the average employment rate among the treated subjects was 62.9%, while the average among the control subjects was 57.6%. The standard deviation of the 59 site ATE estimates is about 7.2 percentage points.

In many sites, assignment to welfare-to-work programs increased the probability of employment within the first two years. However, interpreting these site-specific estimates collectively is difficult because there are several different possible sources of treatment effect variation across sites. For example, a welfare-to-work program might be more effective in urban sites, but it also might be more effective for non-Hispanic White people, who are more common in rural and suburban sites than in urban ones. A principled framework for cross-site treatment effect comparisons is needed to disentangle these sources of variation.

We adopt the framework from Section 4. Under this framework, we first estimate what the ATE in each site would be if all sites had the same population of subjects. We can then attribute the remaining treatment effect variation across sites to differences in site-level or unobserved unit-level covariates. In this case, we transport the ATE estimates to the overall population of 69,399 subjects. In Online Appendix D.3, we also transport the ATE estimates to the population of units in the PI study for further analysis. We transport by solving the optimization problem in Equation 13, where $\phi_\tau(\cdot)$ and $\phi_0(\cdot)$ each map the 23 observed covariates to a basis that includes the original covariates and the interactions between a covariate indicating zero prior earnings and covariates indicating race (see Online Appendix D for more information about these covariates). We standardize the covariates and their interactions to have unit variance before solving Equation 13.

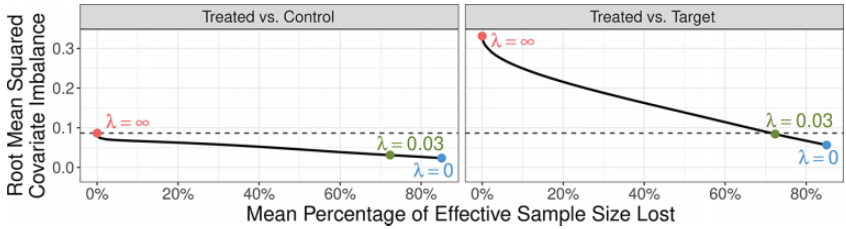

FIGURE 4. *Reductions in covariate imbalance and effective sample size due to weighting with different levels of regularization. The dashed line marks the level of covariate imbalance observed between the treated and control groups without weighting. This imbalance is due to chance randomization alone. Regularization manages the trade-off between covariate imbalance (bias) and effective sample size (variance).*

As discussed in Section 3, we must choose the regularization parameter λ in Equation 13. To do so, we first solve Equation 13 for a range of values of λ and then compare the resulting reductions in covariate imbalance and effective sample size. Given weights $\hat{\gamma}$, we measure covariate imbalance between the treated and control groups by computing within each site the sum of squared differences in the treated and control groups' weighted average covariate values, then averaging over the sites:

$$\text{Treated vs. Control Imbalance} = \frac{1}{59}\sum_{j=1}^{59}\left\| \frac{1}{n_j}\sum_{i:S_i=j}\hat{\gamma}_i \frac{Z_i - \pi_j}{\pi_j(1-\pi_j)}\phi_0(X_i)\right\|_2.$$

We also measure covariate imbalance between the treated group and the target population by computing for each site the sum of squared differences between the treated subjects' weighted average covariate values and the target population's average covariate values, then averaging over the sites:

$$\text{Treated vs. Target Imbalance} = \frac{1}{59}\sum_{j=1}^{59}\left\| \frac{1}{n_j\pi_j}\sum_{i:S_i=j}\hat{\gamma}_i Z_i \phi_\tau(X_i) - \mathbb{E}_{P^*}\{\phi_\tau(X)\}\right\|_2.$$

To measure overall effective sample size, we compute Kish's effective sample size (Kish, 1965) in each site and then average those values over the sites.

Figure 4 shows the trade-off between imbalance reduction and effective sample size reduction. Since treatment was randomized in each site, the treated and control groups were fairly balanced even before weighting. We therefore focus on covariate imbalance between the treated group and the target distribution. Setting $\lambda = 0.03$ reduces covariate imbalance by 80%. The remaining imbalance is less than the imbalance between the unweighted treated and control groups, which, again, is due to chance randomization alone. This imbalance reduction comes at the cost of an approximately 70% reduction in effective sample size. Figure 5 shows in more detail the covariate imbalance in each site before and

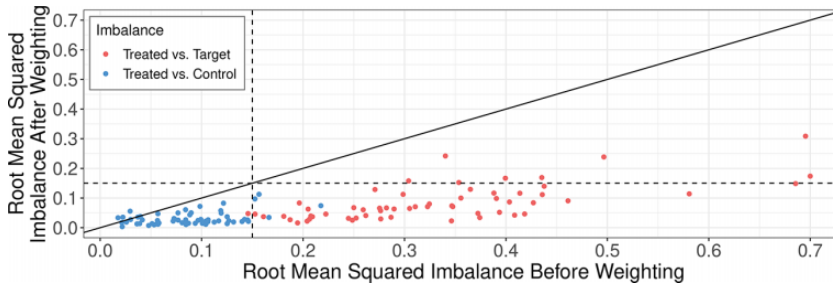

FIGURE 5. *Root mean squared covariate imbalance between the treated group and the control group (blue) and between the treated group and the target population (red) before and after weighting. The covariates are standardized to have unit variance. In most sites, the covariate imbalance between the treated and control groups before weighting (i.e., due to chance randomization alone) did not exceed 0.15, marked by the vertical dashed line. In most sites, weighting brought the covariate imbalance between the treated group and the target population within this range.*

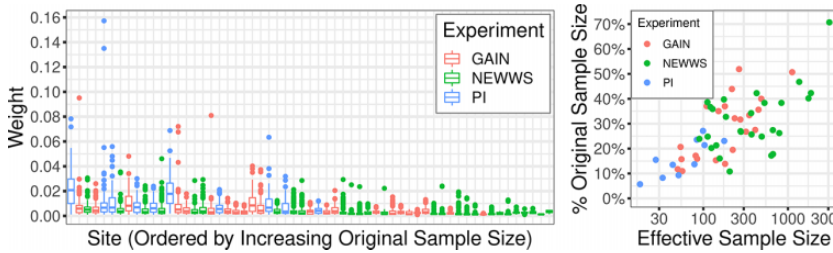

FIGURE 6. *Distribution of nonzero weights in each site (left) and effective sample size after weighting, plotted also as a percentage of the original unweighted sample size (right).*

after weighting with $\lambda = 0.03$. In most sites, the covariate imbalance between the treated and control groups due to chance randomization alone (before weighting) did not exceed 0.15. In only seven sites did the imbalance between the treated group and the target population after weighting exceed that approximate threshold. This indicates that weighting with $\lambda = 0.03$ reduces covariate imbalance well across the board.

To better visualize the cost of transporting the site ATE estimates in terms of sample size, Figure 6 shows the boxplots of the distribution of weights within each site, with weights less than 0.1% excluded. The distribution of weights in some sites is heavily skewed, so that a few subjects are given outsize weight. This lowers those sites' effective sample sizes and increases the standard errors of their transported ATE estimates. Figure 6 also plots the effective sample size of each site as a percentage of the site's original sample size. In most sites, the

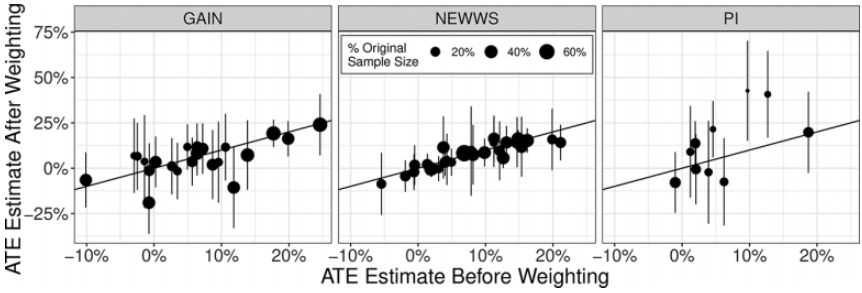

FIGURE 7. *Point estimate and 95% confidence interval for the transported average treatment effect (ATE; in percentage points), plotted against the untransported ATE estimates. Points are sized by percentage effective sample size after weighting.*

reduction in sample size was substantial. The median sample size before weighting was 729, while the median effective sample size after weighting was 181; the smallest effective sample size in a site was 18. Our simulation results highlight the potential problems from such a dramatic drop in effective sample size: Figure 2 shows that there is a point at which the cost (in terms of RMSE) of reducing the effective sample size is greater than the benefit of reducing the bias.

Overall, PI sites had more skewed distributions of weights and lower effective sample sizes than NEWWS and GAIN sites: This motivates our supplemental analysis in Online Appendix D.3 that targets the population of PI subjects instead of the overall population across all three experiments.

Based on these trade-offs, we choose to optimize the weights in Equation 13 with $\lambda = 0.03$. Figure 7 shows the resulting transported ATE point and 95% confidence interval estimate in each site, plotted against the untransported ATE estimates. We use Equation 12 to compute the standard errors of our transported estimates. Transportation substantially changed some sites' ATE estimates, though most sites saw little change.

Transportation also modestly *increased* the estimated overall variability in site ATEs. We estimate that the standard deviation of the true untransported ATEs is $\widehat{\theta}^{\frac{1}{2}} = 0.055$, with a 95% confidence interval of [0.048, 0.075]. The corresponding estimate for the true transported ATEs is $\widehat{\theta}^{*\frac{1}{2}} = 0.060$, with a 95% confidence interval of [0.044, 0.089]. Using our pseudo-R^2 measure, we estimate a negative R^2 of -0.19, indicating a 19% *increase* in the estimated cross-site variance of impacts after adjusting for observed unit-level covariate distributions. This point estimate suggests that differences in the distributional makeup of the different sites could be masking differences in the overall efficacy of treatment implementation at these sites. That being said, the overlapping confidence intervals on both standard deviations suggest that we should take the pseudo-R^2 point estimate as approximate.

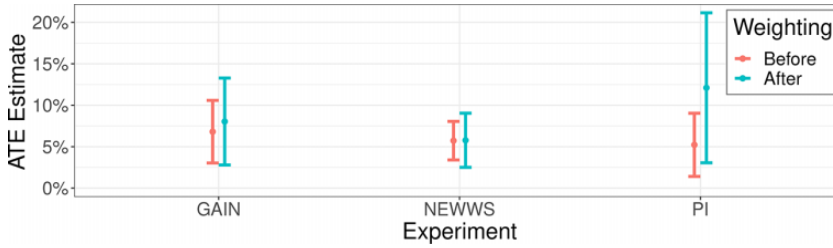

FIGURE 8. *Point estimate and 95% confidence interval for the average treatment effect (in percentage points) in each experiment before and after transportation.*

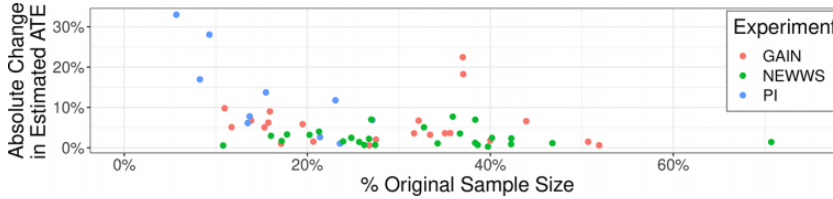

FIGURE 9. *The absolute, percentage-point change in each site's estimated average treatment effect after weighting, plotted against the effective sample size as a percentage of original sample size.*

Obtaining a confidence interval for the R^2 measure itself is complex due to the unknown correlation structure between θ and θ^*; we leave this to future work. Overall, the low magnitude of the R^2 estimate, representing the small shift from $\widehat{\theta}$ to $\widehat{\theta}^*$, and the low magnitudes of $\widehat{\theta}$ and $\widehat{\theta}^*$ themselves suggest that cross-site treatment effect heterogeneity in these experiments is primarily driven by site-level features (or unobserved unit-level features), not by systematic differences in populations, as measured by our observed covariates, across sites. This is consistent with others' findings (e.g., Bloom et al., 2003).

Analyses such as the one presented above could affect conclusions about the relative effectiveness of program types. Figure 8 plots the estimated overall ATE in each experiment before and after transportation. Before transportation, treatment appeared most effective in the GAIN experiment and least effective in the PI experiment. After transporting the sites to the same target distribution of unit-level covariates, we see some evidence that the treatment was least effective in the NEWWS experiment, but we are not confident about this conclusion, given the uncertainties associated with the point estimates. For example, we see in Figure 8 a meaningful increase in the ATE estimate and associated uncertainty for the PI experiment. Figure 9 offers another perspective on this same phenomenon: The weighting process gave substantial weight to only a small percentage

of subjects in the PI experiment, and treatment affected those subjects much differently than it affected other, unrepresentative subjects. The unrepresentative nature of the subjects in the PI experiment compared to the overall population, combined with the fact that PI sites were generally smaller to begin with, as Figure 6 shows, hinders comparisons between the PI experiment and the others. This was perhaps to be expected, given how dramatically different the PI subjects were from the other experiments' subjects with respect to welfare status and prior earnings, as shown in Figure 1. A strong majority of PI subjects were welfare applicants, but very few had received welfare continuously for the prior 12 months, and most had nonzero prior earnings. By contrast, the opposite was true of subjects in most GAIN and NEWWS sites. See Online Appendix D.3 for the results from transporting each site to the distribution of PI units, which in effect assesses each site's performance on this distinct population. We find there that the ATE for NEWWS sites is substantially reduced, further indicating that treatment effect differences across the studies are partially driven by differences in the populations served.

7. Discussion

It is important for researchers and policymakers to understand how treatment effects vary across contexts. Multisite trials can help develop this understanding because the treatments are observed in a variety of settings. However, variation in site ATEs can reflect variation in not only the populations of individuals but also contextual and site-level characteristics. Building on the literature on treatment effect generalization and transportation, we propose to estimate what the site ATEs would be if the sites all had the same population of units. We first develop an approximate balancing-weights procedure that transports estimates from a single site to a given target population and show how the procedure compares to more traditional IPW. We then extend this to the multisite trial setting and show via simulation that this framework is applicable to some very large multisite trials, though power is limited in more typical settings. Applying this method to welfare-to-work experiments that randomly provide some participants access to job training and adult education, we find that heterogeneity in the sites' populations could be masking differences in how effectively the treatment was implemented across sites.

There are several avenues for future work. First, the choice of target population is critical for addressing different substantive questions. In our main analysis, we choose to reweight to the overall population. But we could instead target all participants who were unemployed at baseline, for example. Targeting different populations might lead to substantively different conclusions. If there are important and substantial interactions between unit-level and site-level attributes in determining treatment effects, then we may wish to use several target populations and characterize the variation across the different targets. Future work

might seek to characterize and present this variation in an interpretable and useful way. Alternatively, we could ask a different question: For any given site *j*, how would the ATE in the site change if the site had the *k*th site's distribution of observed unit-level covariates, for $k \in \{1, \ldots, J\}$? This question can be answered by transporting the *j*th site's ATE estimate to each site—that is, by integrating the *j*th site's CATE function over each site's observed covariate distribution. This analysis is conceptually similar to "internal benchmarking" in the criminology literature (Ridgeway & MacDonald, 2014). Clarifying and offering guidance on these choices is an important direction for future research.

Second, we propose to construct weights by decomposing the estimation error principally into (1) error due to imbalance in baseline potential outcomes between the treated and control groups and (2) error due to imbalance in the CATE function between the treated group and the target population. But other decompositions are possible and could lead to different forms of the weighting optimization problem (Equation 11) with different computational and statistical properties. Similarly, throughout this work, we have considered the unit-level covariates of interest to be known and fixed a priori. However, different sets of covariates will lead to different CATE functions and different transported site treatment effects. How to decide *which* covariates are of interest is important for future work (see, e.g., Egami & Hartman, 2021).

Third, the results produced by our method could inform the design of future multisite trials. As we discuss in Section 4.2, there is interest in selecting sites that would best enable generalization of the experimental results to a specific target population (Tipton, 2014; Tipton et al., 2019). If certain unit-level covariates have limited value in predicting cross-site variation, as we find in our application, then researchers can instead focus on balancing other features during site selection. This is especially true when considering the size of the future study.

Finally, we have focused on multisite randomized trials. A natural extension is to multisite settings in which each "trial" is an observational study or quasi-experiment. For instance, estimating teacher value-added modeling has a very similar structure to the multisite trial setting. And, as in our setting, we are often interested in adjusting for (observed) compositional differences across students (Rothstein, 2010). One possibility is to adapt our framework, along with recent results from Chattopadhyay et al. (2022) on generalizing observational studies, to this setting.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received financial support for the research, authorship, and/or publication of this article: This research was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010. This material is based upon work supported by the National Science Foundation under Grant No. 1745640.

Notes

1. We emphasize here that, although the methods used for this analysis are largely the same as those typically used for transportation or generalization, the underlying causal assumptions are different. Unusually, we do not assume that we observe all the unit-level covariates that explain treatment effect heterogeneity because we have a different, more modest goal in mind. The typical goal of generalization or transportation is to understand what the average treatment effect observed in one setting would be in another. By contrast, we seek to understand what the site-level estimates would be after adjusting for observed compositional differences. The former requires that all relevant differences be observed and accounted for. The latter does not.
2. In principle, models $\mathcal{M}$ and $\mathcal{T}$ have separate hyperparameters, represented by the constants C_0 and C_τ. Equation 13 instead has a single tuning parameter, λ, which jointly controls the bias-variance trade-off for both the outcome and treatment models. Including a separate tuning parameter for each model is possible but would come at the cost of greater complexity.
3. Solutions to Equations 13 and 14 can be found using the OSQP solver (Stellato et al., 2020).
4. With the radial basis function kernel, the implied basis representation is the probability density function of a multivariate isotropic Gaussian variable centered at x. The imbalance in this basis will be high if there is little overlap on average between the two Gaussian distributions defined by the individual units in the two populations.

References

Ackerman, B., Schmid, I., Rudolph, K. E., Seamans, M. J., Susukida, R., Mojtabai, R., & Stuart, E. A. (2019). Implementing statistical methods for generalizing randomized trial findings to a target population. *Addictive Behaviors*, *94*, 124–132.

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics*, *130*(3), 1117–1165.

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J., Megicks, S., & Henderson, R. (2017). "Contagious accountability": A global multisite randomized

controlled trial on the effect of police body-worn cameras on citizens' complaints against the police. *Criminal Justice and Behavior*, *44*(2), 293–316.

Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(4), 597–623.

Ben-Michael, E., Feller, A., Hirshberg, D. A., & Zubizarreta, J. (2021). The balancing act for causal inference. *arXiv preprint arXiv, 2110*, 14831.

Ben-Michael, E., Feller, A., & Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, *116*(536), 1789–1803.

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management*, *22*(4), 551–575.

Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness*, *10*(4), 877–902.

Bloom, H. S., Unterman, R., Zhu, P., & Reardon, S. F. (2020). Lessons from New York city's small schools of choice about high school features that promote graduation for disadvantaged students. *Journal of Policy Analysis and Management*, *39*(3), 740–771.

Bloom, H. S., & Weiland, C. (2015). *Quantifying variation in head start effects on young children's cognitive and socio-emotional skills using data from the national head start impact study*. Technical report, Manpower Demonstration Research Corporation.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, *97*(1), 65–108.

Chattopadhyay, A., Cohn, E. R., & Zubizarreta, J. R. (2022). One-step weighting to generalize and transport treatment effect estimates to a target population. *arXiv preprint arXiv:2203.08701*.

Chattopadhyay, A., Hase, C. H., & Zubizarreta, J. R. (2020). Balancing vs modeling approaches to weighting in practice. *Statistics in Medicine*, *39*(24), 227–3254.

Crepon, B., Duflo, E., Huillery, E., Pariente, W., Seban, J., & Veillon, P-A. (2018). *Cream skimming and the comparison between social interventions: Evidence from entrepreneurship programs for at-risk youth in France*. Mimeo.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199.

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, *221*(2), 644–654.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., & Hernán, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, *39*(14), 1999–2014.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., & Hernán, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, *75*(2), 685–694.

Dehejia, R. H. (2003). Was there a Riverside miracle? A hierarchical framework for evaluating programs with grouped data. *Journal of Business & Economic Statistics*, *21*(1), 1–11.

Ding, P., Feller, A., & Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*, *114*(525), 304–317.

Djebbari, H., & Smith, J. (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, *145*(1-2), 64–80.

Egami, N., & Hartman, E. (2021). Covariate selection for generalizing experimental results: Application to large-scale development program in Uganda. *Journal of the Royal Statistical Society (Series A)*, *184*(4), 1524–1548. https://doi.org/10.1111/rssa.12734

Egami, N., & Hartman, E. (2020). *Elements of external validity: Framework, design, and analysis*. Cambridge University Press & Assessment.

Flores, C. A., & Mitnik, O. A. (2013). Comparing treatments across labor markets: An assessment of nonexperimental multiple-treatment strategies. *Review of Economics and Statistics*, *95*(5), 1691–1707.

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, *15*(3), 965–1056.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25–46.

Harvey, R. A., Hayden, J. D., Kamble, P. S., Bouchard, J. R., & Huang, J. C. (2017). A comparison of entropy balance and probability weighting methods to generalize observational cohorts to a population: A simulation and empirical example. *Pharmacoepidemiology and Drug Safety*, *26*, 368–377.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217.

Hirshberg, D. A., Maleki, A., & Zubizarreta, J. R. (2019). Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*.

Hirshberg, D. A., & Wager, S. (2021). Augmented minimax linear estimation. *Ann. Statist*. *49*(6): 3206–3227 DOI: 10.1214/21-AOS2080

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, *125*(1–2), 241–270.

Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, *50*(2), 373–419.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Josey, K. P., Berkowitz, S. A., Ghosh, D., & Raghavan, S. (2021). Transporting experimental results with entropy balancing. *Statistics in Medicine*, *40*(19): 4310–4326. https://doi.org/10.1002/sim.9031

Josey, K. P., Yang, F., Ghosh, D., & Raghavan, S. (2022). A calibration approach to transportability with observational data. *Statistics in Medicine*. *41*(23): 4511–4531. https://doi.org/10.1002/sim.9523

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539.

Keele, L., Ben-Michael, E., Feller, A., Kelz, R., & Miratrix, L. (2020). Hospital quality risk standardization via approximate balancing weights. *arXiv preprint arXiv:2007.09056*.

Kemple, J. J., & Haimson, J. (1994). *Florida's project independence. Program implementation, participation patterns, and first-year impacts*. Technical report, Manpower Demonstration Research Corporation.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103–127.

King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, *14*(2), 131–159.

Kish, L. (1965). *Survey sampling*. John Wiley & Sons.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4156–4165.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, *7*(1), 295–318.

Ma, X., & Wang, J. (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, *115*(532), 1851–1860.

May, H., Goldsworthy, H., Armijo, M., Gray, A. M., Sirinides, P. M., Blalock, T. J., Anderson-Clark, H., Schiera, A. J., Blackman, H., & Gillespie, J. (2014). *Evaluation of the i3 scale-up of reading recovery year two report, 2012–13*. Technical report, Consortium for Policy Research in Education.

Miratrix, L. W., Sekhon, J. S., & Yu, B. (2011). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(2), 369–396.

Miratrix, L. W., Weiss, M. J., & Henderson, B. (2021). An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness*, *14*(1), 270–308.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*(4), 465–472. (Original work published 1923)

Nie, X., Imbens, G., & Wager, S. (2021). Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., Jenkins, F., Fletcher, P., Quinn, L., Friedman, J., Ciarico, J., Rohacek, M., Adams, G., & Spier, E. (2010). Head Start impact study. Final report. *Administration for Children & Families*.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 99.

Raudenbush, S. W., Reardon, S. F., & Nomi, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of Research on Educational Effectiveness*, *5*(3), 303–332.

Riccio, J., & Friedlander, D. (1992). *GAIN: Program strategies, participation patterns, and first-year impacts in six counties. California's greater avenues for independence program*. Technical report, Manpower Demonstration Research Corporation.

Ridgeway, G., & MacDonald, J. M. (2014). A method for internal benchmarking of criminal justice system performance. *Crime & Delinquency*, *60*(1), 145–162.

Robins, J., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, *22*(4), 544–559.

Rosenbaum, P. R. (2009). *Design of observational studies*. Springer Science & Business Media.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214.

Rudolph, K. E., Schmidt, N. M., Glymour, M. M., Crowder, R., Galin, J., Ahern, J., & Osypuk, T. L. (2018). Composition or context: Using transportability to understand drivers of site differences in a large-scale housing experiment. *Epidemiology*, *29*(2), 199–206.

Rudolph, K. E., & van der Laan, M. J. (2017). Robust estimation of encouragement-design intervention effects transported across sites. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(5), 1509–1525.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods*. National Center for Education Evaluation and Regional Assistance.

Stanton, A. L., Ganz, P. A., Kwan, L., Meyerowitz, B. E., Bower, J. E., Krupnick, J. L., Rowland, J. H., Leedham, B., & Belin, T. R. (2005). Outcomes from the moving beyond cancer psychoeducational, randomized, controlled trial with breast cancer patients. *Journal of Clinical Oncology*, *23*(25), 6009–6018.

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., & Boyd, S. (2020). OSQP: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, *12*(4), 637–672.

Teo, T. (2014). *Handbook of quantitative methods for educational research*. Springer Science & Business Media.

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501.

Tipton, E., Yeager, D. S., Iachan, R., & Schneider, B. (2019). Designing probability samples to study treatment effect heterogeneity. In P. J. Lavrakas, M. W. Traugott, C. Kennedy, A. L. Holbrook, E. D. de Leeuw, & B. T. West (Eds.). *Experimental methods in survey research: Techniques that combine random sampling with random assignment* (pp. 435–456). Wiley.

Walters, C. R. (2015). Inputs in the production of early childhood human capital: Evidence from head start. *American Economic Journal: Applied Economics*, *7*(4), 76–102.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across

sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., & Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, *186*(8), 1010–1014.

Yang, S., & Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, *105*(2), 487–493.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, *110*(511), 910–922.

Authors

Benjamin Lu is a Ph.D. candidate at UC Berkeley, Evans Hall, Berkeley, CA 94720; e-mail: b.lu@berkeley.edu. His research interests lie at the intersection of statistics and law.

Eli Ben-Michael is an Assistant Professor at Carnegie Mellon University, 4800 Forbes Ave, Pittsburgh, PA 15213; e-mail: ebenmichael@cmu.edu. His research focuses on developing statistical and computational methods to solve practical issues in public policy and social science research.

Avi Feller is an Associate Professor at UC Berkeley, Goldman School of Public Policy and Department of Statistics, 2607 Hearst Ave, Berkeley, CA 94720; e-mail: afeller@berkeley.edu. His research focuses on the interface of statistics and data science with the social sciences.

Luke Miratrix is an Associate Professor at the Harvard Graduate School of Education, 14 Appian Way, Cambridge, MA 02140; e-mail: lmiratrix@g.harvard.edu. His research focuses on developing and elucidating statistical methods for causal inference in the social sciences, with an emphasis on methods for randomized trials.

Manuscript received March 26, 2021
Revision received December 6, 2022
Accepted December 15, 2022